RANDOM NUMBER GENERATION ON THE CDC 1604

JOSEPH M. BARRON

[redacted]

RANDOM NUMBER GENERATION

ON THE CDC 1604

* * * * *

Joseph M. Barron

RANDOM NUMBER GENERATION ON THE CDC 1604

by

Joseph M. Barron

Lieutenant Commander, United States Navy

Submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

United States Naval Postgraduate School
Monterey, California

1 9 6 2

[redacted]

RANDOM NUMBER GENERATION

ON THE CDC 1604

by

Joseph M. Barron

This work is accepted as fulfilling

the thesis requirements for the degree of

MASTER OF SCIENCE

from the

United States Naval Postgraduate School

ABSTRACT

A brief history of the development of tables of random numbers is presented. This is followed by an investigation of the generation of uniformly distributed random numbers on the CDC 1604 digital computer. Certain congruential relations are examined in order to determine parameters that will produce acceptable random number sequences. Statistical tests of the generated numbers are described and the results of these tests are presented.

PREFACE

Statistical sampling techniques using a "Monte Carlo" method can require a large supply of random numbers. Such techniques have been used in the evaluation of definite integrals, the solution of ordinary and partial differential equations, the solution of integral equations and matrix inversion, to mention only a few applications. On the other hand, some problems, described in terms of a stochastic model, can only be solved by recourse to sampling techniques, as for example, mathematical war gaming problems.

When such models are programmed for a high speed digital computer, it is necessary to produce large numbers of random numbers quickly. The tables of random digits or random uniform numbers customarily used in hand calculations do not lend themselves to use in the high speed computer. It is inefficient to read into the computer the large number of random digits required to meet the demand for such numbers arising in problems which are otherwise within the capabilities of the machine. Using the same numbers repeatedly in a given problem would introduce bias.

It has been suggested that special equipment which utilizes a physical random process be used to produce the random numbers. On the surface, this would seem desirable but, in practice, difficulties immediately arise. First, the construction and maintenance of such a device would be quite expensive. To be useful, it would be required to produce a number in the order of a few microseconds for extended periods. Reliability requirements would necessitate its output to be continually checked for randomness. The second difficulty is that such a random device would not permit the recalculation of a problem using the same generated numbers. This could prove to be a handicap in checking on machine operations.

These difficulties lead to the exploration of arithmetical schemes for generating numbers which take advantage of the computer's high speed and place minimum demands on computer memory storage. Such numbers could not be considered truly random; hence, the adoption of the term "pseudo-random numbers".

It is the intent of this paper to present a brief historical development of random number generation and to explore, in particular, the multiplicative congruential type of generator for adaptation to the CDC 1604.

TABLE OF CONTENTS

TABLE OF SYMBOLS

mod M	The abbreviation for modulus m.
$\equiv$	The number theory symbol meaning, "is congruent to".
X_v^2	A chi-square random variable with v degrees of freedom.
G_v	Distribution function of a chi-square random variable with v degrees of freedom, i.e. $P\left[X_v^2 \leq x\right] = G_v(x)$
$X_{v:a}^2$	The real number defined by $1 - G_v\ (X_{v:a}^2) = a.$
Uniform $[0-1]$	The uniform distribution over the unit interval.

I. INTRODUCTION

1. Kendall and Smith's Random Sampling Numbers.

It might be supposed that if one wanted to obtain a sequence of random digits that almost any method that employed some sort of chaotic selection would suffice. Observing the indications of a roulette wheel or, perhaps, an unsystematic selection of digits from a large telephone directory might seem like "reasonable" methods. Once the digits were collected, however, a problem would arise. By what criterion could their randomness be judged?

It was to this problem that M. G. Kendall and B. Babington Smith addressed themselves in their now classic work, "Randomness and Random Sampling Numbers". [1] [1] They start their paper with a discussion on the relation between randomness and probability.

> In colloquial speech the word "random" is applied to any method of choice which lacks aim or purpose; and this usage is also found in certain sciences. In statistics, however, the word has a somewhat different and more definite significance, closely related to probability. It appears, in fact, that for statistical purposes the ideas of randomness and probability are inseparable, whether one belongs to the "intuitive" school which regards probability as an undefinable, or to the opposing "frequency" school which seeks to explain it in terms of statistical frequencies.

Having linked the concepts of probability and randomness, Kendall and Smith concede that there are many controversial areas that pose abstract problems that are,--- "verging, at times, on the theological." In the interest of getting on with the more mundane problems of Random Sampling Numbers, they state:

> We take randomness and probability to be undefined ideas obeying certain intuitively formulated principles, which will be found sufficient to give results of practical application.

[1] Numbers in brackets refer to references cited in the Bibliography at the end of the paper.

The authors note that any combination of digits could be considered a random set, in that all combinations are possible when drawing from an infinite universe of digits. However, not all combinations are practical for random sampling; for example, a block of digits which are all the same. For purposes of sampling, a definition of Random Sampling Number is made: a set of numbers which can be used for random sampling, not necessarily a set obtained by random methods. It is Kendall and Smith's intention to produce a table of random digits and to test these for acceptability as Random Sampling Numbers. That is, Random Sampling Numbers is a collective term for a series of random digits.

Such a set of Random Sampling Numbers should conform to the notion of relative frequency. That is, if samples were taken from a universe consisting of an infinite number of digits such that each of the ten digits 0 through 9 may be expected to appear with the same frequency, then, it is expected that in a large number of trials each of the digits will appear approximately an equal number of times. Further, it would be expected that each pair of digits would occur approximately the same number of times, similarly triples, etc.; that is, no digit would tend to follow another digit in any consistent pattern.

Here then is an important distinction: any given set of N numbers need not follow such expectations but a set of Random Sampling Numbers should do so. Hereafter, we shall identify such a set of Random Sampling Numbers as a random set.

In order to ascertain whether a given set of digits is a good enough "approximation" to a random set, the observed frequencies may be compared to the theoretical frequencies by means of a chi-square test. On "purely arbitrary grounds", Kendall and Smith elected to say that a set of numbers is acceptable if the resulting chi-square statistic is within the range of $0.01 \leq P \leq 0.99$, where, with O_j the observed and E_j the expected frequencies of the $_j$th digit,

$$X^2 = \sum_{j=o}^{9} \frac{(O_j - E_j)^2}{E_j}$$

and

$$P = \frac{\int_{x}^{\infty} e^{-\frac{X^2}{2}} X^{9-1} \, dX}{\int_{o}^{\infty} e^{-\frac{X^2}{2}} X^{9-1} \, dX}$$

with 9 being the number of degrees of freedom.

Having thus determined the characteristics of this set, the authors then proceed to establish a series of four tests for a set of digits to be considered as a random set. Subsequently, these tests have become standard in the literature on random numbers. Briefly, they are as follows:

(a) The frequency test to determine if all the digits of a set occur in approximately equal amounts.

(b) The serial test which tests to see if any digit tends to follow another.

(c) The poker test which is named after the card game and examines the digits in blocks of five for five of a kind, four of a kind etc. The frequency of these hands are then compared with the expected distribution of such hands.

(d) The gap test which inquires into the gaps occurring between the same digits in a series and compares this with the theoretical distribution.

Although these tests are not sufficient, the authors feel that collectively they are powerful.

With the foregoing criteria established, Kendall and Smith attempt to set up a random series by selecting digits from the London Telephone Directory in a haphazard fashion. Even though they took what would seem like extra precautions to avoid bias, the results of their tests indicated significant bias.

They then constructed an electro-mechanical device to produce random digits. Essentially, it consisted of a wheel divided into ten equal segments with a digit assigned to each segment. It was driven by an electric motor whose speed was accurately controlled. The operator on a separate circuit controlled a neon lamp whose flash was of such a short duration that the moving wheel appeared to be stationary. The number that appeared under a previously fixed pointer was recorded as the random digit. The actuation of the lamp was kept haphazard by having the operator move a stylus over an intricate maze of circuitry which, it was maintained, caused the light to operate in a random fashion.

A table of 100,000 digits was produced in this fashion and, based on their tests, were considered acceptable as Random Sampling Numbers [1].

2. RAND Corporation's Table of Random Digits.

Kendall and Smith's table appears to have been satisfactory for a number of years; at least, there were no other tables published. In 1947, a group of research workers in the RAND Corporation responded to a growing need for random numbers by producing a table of one million random digits [2] . These were needed to solve various computerized research problems by experimental probability procedures. Such procedures were called by the colorful name, Monte Carlo Methods. For many of the problems that RAND was working on, the need for numbers far exceeded those of Kendall and Smith's Tables. In fact, it meant that these tables would have to be used time and again in a particular problem. This presented the consequent danger of introducing unwanted correlations. Such a growing thirst for random numbers was, of course, due to new problems and rapidly developing computer technology available for solving these problems.

RAND'S table was produced by electronic-mechanical means and was apparently more sophisticated than Kendall and Smith's machine.

> In principle the machine was a 32-place roulette wheel which made, on the average, about 3,000 revolutions per trial and produced one number per second.

It appears that the engineers designing the device did have many problems because the original machine showed statistically significant biases and had to be modified extensively. Further, even though electronic checks were continually made, after a month the results of tests of a block of numbers produced showed significant bias. This indicated that the machine was running down even though the electronic checks were acceptable.

3. Pseudo-random Number Generation.

While RAND'S table was suitable for work with punched card computers, it proved impractical with the advent of high speed electronic computers because:

(a) it placed severe storage requirements on the machine

(b) reading in the tables was slow

(c) much larger tables were required for the solution of certain types of problems.

These shortcomings led to an interest in methods for producing "pseudo-random" numbers by arithmetic means. A definition of this term was given by D. H. Lehmer as:

> A pseudo-random sequence is a vague notion embodying the idea of a sequence in which each term is unpredictable to the uninitiated and whose digits pass a certain number of tests traditional with statisticians and depending somewhat on the uses to which the sequence is to be put.

One of the first methods used was proposed by von Neumann and Metropolis wherein an arbitrary n-digit number, Y_0, was selected. This number was then squared and Y_1 was produced as an n-digit number from the center of the 2 n-digits of Y_0^2. The process was then repeated, producing Y_{i+1} from Y_i.

Although this process generated pseudo-random numbers, its period was not predictable. Further, it was demonstrated by D. H. Lehmer [3] that, in general, it would not produce a very long cycle before the process degenerated into a sequence of zeros. Consequently, it could not be recommended as a source of great quantities of random numbers.

During the second symposium at the Harvard Computation Laboratory, D. H. Lehmer suggested [3] that the multiplicative congruential relation (See Appendix A)

$$Y_{r+1} \equiv L\, Y_r \pmod{M} \qquad (1\text{-}1)$$

could be used where the least positive residue is taken as the generated number. In effect, this statement says that $(Y_{r+1} - LY_r) / M$ is an integer. The modulus M could be selected so as to be compatible with the base of the machine being used; that is, it could be of the form (2^P) or (10^P). P normally is selected to be the number of positions available in a word of the computer being used. Further, in a binary machine, with a proper selection of L, the period of this generator is $2^{(P-2)}$.

Using P in this way avoids the necessity for performing the final division operation in most high speed computers. For example, in the CDC 1604 using a modulus of 2^{47}, the multiplication (MUI) of the number L by Y_r positions the product in the QA registers. The 47 low-order bits of this product in the A register are, in fact, the desired residual or the newly generated random number.

Although such a method is relatively fast, nevertheless, it does use the multiplication command which, for a computer, is time consuming. Greenberger [4] proposed that L in equation (1-1) be modified to

$$Y_{r+1} \equiv (2^A + 3) Y_r \qquad (\text{mod } 2^P) \qquad (1\text{-}2)$$

with $A \geq 3$. By doing this, the multiplication could be performed by a shift command and 3 add commands which represented a considerable saving in time over the straight multiplication instruction. Further, he shows [5] that an L of this form does satisfy the conditions necessary to insure the full period of 2^{P-2}.

Another variation of equation (1-1) was proposed by Rotenberg [6] whose generator became

$$Y_{r+1} \quad (2^A + 1) Y_r + C \qquad (\text{mod } 2^P) \qquad (1\text{-}3)$$

This again was faster than Greenberger's model by one add instruction but, unlike the former, this gave a full period of 2^P for $A \geq 2$ and C an odd integer.

Equations (1-2) and (1-3) will be investigated in Chapters II and III for use as pseudo-random number generators for the CDC 1604 computer.

II. EMPIRICAL TEST PROCEDURES

1. General

Previous studies made using other machines [5] [6] [7] indicated that the quality of the generated pseudo-random numbers using equations (1-2) and (1-3) can be very sensitive to the selection of A. Aside from a guarantee of the period length (See Appendix A), there is, presently, little mathematical theory to aid in the selection of parameters that will enhance the random attributes of the generated samples. The task is largely an empirical one wherein an A and a starting value, Y_o, are selected and the subsequent output is submitted to a battery of tests. Such empirical analysis can only give general indications as to desirable parameters because it is only practical to investigate a small portion of the total cycle of the generator. This fact can be appreciated by observing that using a modulus of 2^{47} in (1-2) the fraction of the full period that would be considered, if even, say 10^9 elements were tested, is approximately 10^{-3}.

As has been cited in Chapter I, the tests for randomness offered by Kendall and Smith were designed to check on the randomness of the individual digits of their table of random numbers. The generators under discussion here, however, produce pseudo-random numbers. Consequently, those tests that lend themselves to investigating certain properties associated with random numbers rather than those associated with individual digits are favored in the investigative work herein reported. Should a requirement arise for random digits when using equations (1-2) and (1-3), only the high order (left-most) bit positions of the binary representation of the generated number should be considered. (See Appendix A).

2. Uniformly Distributed Random Variables.

Many uses of computerized sequences of random numbers require that they be uniform $[0,1]$. Not only is this distribution used in its own right but numbers from this distribution may be used to produce numbers with other distributions. Converting the generated pseudo-random numbers of equations (1-2) and (1-3) to uniform $[0,1]$ is quickly accomplished in a high speed digital computer as it merely requires the positioning of a decimal point to the left of the generated numbers. (See Appendix B.)

3. Description of Tests.

The following tests were performed on the blocks of numbers forming a generated sequence.

(a) Frequency test: The classic chi-square goodness of fit test was used as a frequency test [9]. This test divides the unit interval into k equal parts. The generated numbers were each checked and a tally kept on their frequency distribution within the k intervals. This is, perhaps, the most frequently applied of the random tests.

Attempts have been made to establish some sort of an optimal size for k. Mann and Wald, in their paper [8], present a procedure by which the lengths of the class intervals are determined so that the probability of each class under the null hypothesis is equal to 1/k where k is the number of class intervals. They establish a Theorem which says that, in the limit, the best (defined by them) value of k is given by

$$k = 4\left(\frac{2(N-1)^2}{C^2}\right)^{1/5}$$

where C is determined so that

$$\frac{1}{2\pi}\int_{c}^{\infty} e^{\frac{-x^2}{2}}\,d\,x$$

is equal to the size of the critical region (probability of the critical region under the null hypothesis) and N is the number of random elements generated. For significance levels of .01 and .05 the corresponding values of C are 2.327 and 1.645 respectively. These, in turn, lead to $k \approx 88$ for .01 and $k \approx 114$ for .05 with N 5000. A $k = 100$ was used in the large sample tests reported on in Chapter III.

(b) Serial Correlation: This test was set up following the formula given by Kendall [9] . Lags from 1 through 10 were examined.

(c) Run test: The number of runs up and down were examined and compared with the expected theoretical distribution using a chi-square test. The first number to be examined is compared with the next number in the sequence. If the second is larger than the first, a run up of size one has occurred but is not yet recorded. The second and third numbers are then examined in the same way. If the third is greater than the second, a run of size two has occurred but is not yet recorded. If the third is smaller than the second, a run of size one is recorded etc. [10] .

(d) Moments: The first four central moments were computed and compared with those for the uniform $[0,1]$ distribution.

(e) Poker Test: This test was used to examine the sequence of digits within a number. Starting at the highest order bit position, the digits were grouped to give five octal digits (15 bit positions). The poker hand value of these five digits was tallied. The next 15 bit positions of the same random number were similarly examined and tallied. The succeeding generated numbers were examined in a similar fashion. A chi-square test was performed on the results.

4. Small Sample Tests.

Exploratory tests were made on generators (1-2) and (1-3) using two different starting values, $Y_0 = 1$ and $Y_0 = 2^{46} - 1$. Sample sizes of 100 were taken at each: A = 5,10,15....,45. This was done in order to ascertain, in a general way, the sensitivity of the generators to changes in A. The results presented in Tables 1 - 8 of Chapter III indicate that for these particular starting values the selection of A is critical. It further confirmed that a good selection of A was approximately $(2^{47})^{\frac{1}{2}}$ for both generators. This supp rted the rule of thumb that was offered by Greenberger [5] .

Selection of $Y_0 = 1$ also gives an indication of how the generator would respond when faced with a low number. Cases have been reported [7] where, when this occurs, an unfortunate selection of A will make the generator extremely sluggish and can produce several hundred low numbers before yielding a more chaotic output. This succession of low numbers gives, in effect, a sequence all below 0.5. Such a characteristic, obviously, is undesired.

Parallel procedures were followed using a sample size of 1000 numbers at each A.

5. Large Sample Tests.

Having selected a value of A from the results of the small sample tests, five blocks of 10,000 numbers were generated and subjected to the battery of tests.

6. Modifications.

Experiments were conducted to investigate the quality of pseudorandom numbers generated by addressing the output of one generator to a second one having a different value of A. This linking process for equation (1-2) and (1-3) can be described as follows:

$$Y_r \equiv (2^A + 3)\ Y_{r-1} \qquad (\text{mod } 2^{47})$$

$$Q_r \equiv (2^B + 1)\ Y_r + 1 \qquad (\text{mod } 2^{47}) \qquad (2\text{-}1)$$

where $Q_r \left| r = 1, 2, \ldots\ldots\ldots \right|$ becomes the sequence of interest.

III. RESULTS AND CONCLUSIONS

1. Interpretation of Data.

No attempt was made to select the best parameter A based on the small sample tests discussed in Chapter II. Rather, the observed data was prepared in tabular form for each of the generators. This allows general comparisons to be made on the effects of changing the inputs on each of the generators being investigated. In this way, certain ranges of parameters are observed to produce small samples whose performance is consistently poor under the battery of tests. It also permits a comparison of one generator with the other as A is varied. For if one assumes, as will be done in this paper, that the generator that exhibits the least number of these "poor" indications is the preferred generator such a tabulation will make it apparent. This assumption does not take into account the differences in speed of generation between (1-2) and (1-3). In the CDC 1604, this difference is one add instruction or 7.2 micro-seconds per word and although this is a consideration, for purposes of this analysis, it will be considered minor. The same can be said of the differences in cycle length between the generators under consideration because each of them is extremely large. In short, the point of focus in the discussion to follow is to be centered on the generators ability to produce pseudo-random numbers that aré of practical use.

A particular parameter was considered "poor" if it failed to pass anyone of the following arbitrarily selected criteria:

(a) The frequency chi-square statistic, X_9^2 must be $3.3 \leq X_9^2 \leq 16.9$ that is, for 9 degrees of freedom the critical values of the chi-square were selected as .05 and .95.

(b) No long runs; where a long run will be considered a run of length six or greater for samples of size 100 and of length seven or greater in samples of size 1000.

(c) Serial correlation coefficients of lag one must be less than 0.2 for samples of size 100 and 0.05 for samples of size 1000.

(d) Mean greater than or equal to 0.430 and less than or equal to 0.570.

2. Results of Small Sample Tests.

Tables 1 through 8 indicate the results of varying A in the generators

$$Y_{r+1} \equiv (2^A + 3) Y_r \pmod{2^{47}} \tag{3-1}$$

and

$$Y_{r+1} \equiv (2^A + 1) Y_r + 1 \pmod{2^{47}} \tag{3-2}$$

In tables 1 through 4 $Y_0 = 1$ while in tables 5 through 8 $Y_0 = 2^{46} - 1$.

TABLE 1.

TEST RESULTS ON:

$$Y_{r+1} \equiv (2^A + 3)Y_r \quad (\mathrm{mod}\ 2^{47})$$

$$Y_0 = 1$$

TEST		A									THEO. VALUE
		5	10	12	20	25	30	35	40	45	
FREQUENCY X_9^2	1	9.2	4.8	10.8	11.6	13.6	10.8	7.8	16.2	17.4	$X^2_{9:05}$ = 16.919
	2	15.5	4.6	10.4	7.1	4.4	8.5	8.4	4.5	5.9	
SERIAL CORRELATION	1	.132	.049	-.175	.215	.313	.037	.143	-.072	-.150	0.000
	2	.005	.030	-.042	.039	.026	.044	.013	-.032	.018	
MEAN	1	.475	.469	.498	.459	.425	.475	.467	.517	.479	0.500
	2	.488	.500	.508	.488	.493	.498	.490	.500	.502	
VARIANCE	1	.098	.084	.098	.074	.090	.101	.090	.093	.069	0.0833
	2	.086	.083	.082	.087	.082	.084	.082	.084	.083	
RUNS 6	1	1	0	0	0	2	1	0	0	0	0.0285
RUNS 7	2	0	0	0	0	1	1	0	0	1	0.0388

1. Sample size 100
2. Sample size 1000

TABLE 2.

TEST RESULTS ON:

$$Y_{r+1} \equiv (2^A + 3)\, Y_r \quad (\mathrm{mod}\ 2^{47})$$

$$Y_o = 1$$

TEST		A									THEO. VALUE
		16	17	18	19	20	21	22	23	24	
FREQUENCY X^2_9	1	15.4	11.4	8.0	8.0	11.6	9.0	16.8	11.2	15.4	$X^2_{9:05}$ = 16.919
	2	12.6	7.6	5.4	8.3	7.1	7.7	8.8	16.3	14.0	
SERIAL CORRELATION	1	.088	.231	.135	.156	.215	.024	-.006	-.002	.189	0.000
	2	-.040	-.006	.006	.036	.039	-.005	-.038	.019	.091	
MEAN	1	.480	.475	.499	.485	.459	.544	.456	.426	.425	0.500
	2	.523	.492	.502	.502	.488	.504	.495	.513	.471	
VARIANCE	1	.109	.098	.094	.086	.074	.083	.093	.072	.090	0.0833
	2	.084	.086	.084	.082	.087	.086	.084	.087	.084	
RUNS ≥6	1	1	1	0	0	0	0	0	0	1	0.0285
RUNS ≥7	2	1	0	0	0	0	0	0	0	1	0.0388

1. Sample size 100
2. Sample size 1000

TABLE 3.

TEST RESULTS ON:

$$Y_{r+1} \equiv (2^A + 1) Y_r + 1 \pmod{2^{47}}$$

$$Y_0 = 1$$

TEST		A									THEO. VALUE
		5	10	15	20	25	30	35	40	45	
FREQUENCY X^2_9	1	27.0	10.6	23.8	9.0	>50	>50	>50	8.8	>50	$X^2_{9:05}$ = 16.919
	2	4.6	10.1	7.2	5.0	>50	>50	8.1	11.0	>50	
SERIAL CORRELATION	1	.200	.234	.150	.055	.999	.999	.943	.029	.426	0.000
	2	.081	.030	.002	.046	.999	.982	.387	-.018	-.428	
MEAN	1	.410	.483	.479	.467	.409	-.013	.319	.474	.427	0.500
	2	.487	.502	.497	.497	.398	.395	.481	.499	.425	
VARIANCE	1	.090	-.090	.098	.071	.000	.000	.086	.088	.096	0.0833
	2	.085	.086	.082	.081	.001	.087	.085	.081	.100	
RUNS ≥6	1	1	1	1	1	>15	>15	>15	3	0	0.0285
RUNS ≥7	2	1	1	0	1	>15	>15	>15	>15	0	0.0388

1. Sample size 100
2. Sample size 1000

TABLE 4.

TEST RESULTS ON:

$$Y_{r+1} \equiv (2^{A} + 1) Y_r + 1 \pmod{2^{47}}$$

$$Y_0 = 1$$

TEST		A									THEO. VALUE
		16	17	18	19	20	21	22	23	24	
FREQUENCY X^{2}_{9}	1	61	30	16.4	4.8	9.0	4.4	29.6	488	900	$X^{2}_{9:05} =$ 16.919
	2	15.4	7.3	8.8	8.5	5.1	14.8	26.5	4915	900	
SERIAL CORRELATION	1	.765	.308	.185	.204	.055	-.059	-.055	-.404	.99	0.00
	2	.086	.015	-.013	.021	.046	.027	-.011	-.398	.99	
MEAN	1	.348	.385	.417	.467	.467	.468	.390	.145	.204	0.500
	2	.489	.499	.501	.491	.497	.471	.415	.154	.199	
VARIANCE	1	.095	.092	.091	.094	.071	.087	.074	.051	0.00	0.0833
	2	.086	.084	.086	.087	.081	.081	.081	.051	0.00	
RUNS ≥ 6	1	4	1	1	1	1	0	0	0	> 15	0.0285
RUNS ≥ 7	2	> 15	5	0	1	1	0	0	0	> 15	0.0388

1. Sample size 100
2. Sample size 1000

TABLE 5.

TEST RESULTS ON:

$$Y_{r+1} \equiv (2^A + 3)\, Y_r \pmod{2^{47}}$$

$$Y_0 = 2^{46}-1$$

TEST		A									THEO. VALUE
		5	10	15	20	25	30	35	40	45	
FREQUENCY X^2_9	1	17.2	2.6	6.4	19.0	12.4	20.2	5.8	8.4	11.4	$X^2_{9:05}$ = 16.919
	2	4.38	4.15	18.5	17.8	6.9	10.3	7.16	8.5	8.4	
SERIAL CORRELATION	1	0.103	-.150	-.077	.029	-.046	-.007	-.041	.015	.040	0.00
	2	.032	-.009	-.062	-.025	.039	-.010	-.039	.022	-.002	
MEAN	1	.469	.480	.513	.527	.537	.520	.514	.487	.507	0.500
	2	.506	.499	.495	.514	.493	.520	.497	.488	.509	
VARIANCE	1	.075	.077	.073	.077	.084	.075	.075	.085	.076	0.0833
	2	.086	.085	.083	.081	.085	.081	.084	.088	.089	
RUNS ≥ 6	1	1	0	1	0	0	0	0	0	0	0.0285
RUNS ≥ 7	2	0	0	0	0	0	0	0	0	0	0.0388

1. Sample size 100
2. Sample size 1000

TABLE 6.

TEST RESULTS ON:

$$Y_{r+1} \equiv (2^{A} + 3)\, Y_r \quad (\text{mod } 2^{47})$$

$$Y_o = 2^{46} - 1$$

TEST		A									THEO. VALUE
		16	17	18	19	20	21	22	23	24	
FREQUENCY X^2_9	1	10.4	19.4	19.2	9.2	19.0	12.6	7.2	8.0	5.4	$X^2_{9:05}$ = 16.919
	2	5.0	9.1	6.7	13.8	17.8	5.7	10.6	9.0	19.1	
SERIAL CORRELATION	1	-.181	.063	-.102	-.030	.029	.097	.021	-.118	-.014	0.00
	2	.025	-.016	.022	-.036	-.025	-.022	.023	-.033	.046	
MEAN	1	.513	.429	.438	.511	.527	.443	.504	.500	.455	0.500
	2	.507	.509	.491	.514	.514	.486	.487	.507	.481	
VARIANCE	1	.073	.068	.081	.071	.077	.073	.072	.096	.081	0.0833
	2	.083	.086	.085	.081	.081	.083	.082	.086	.086	
RUNS ≥ 6	1	0	0	0	0	0	0	0	1	0	.0285
RUNS ≥ 7	2	0	0	0	0	0	0	0	1	0	.0388

1. Sample size 100
2. Sample size 1000

TABLE 7.

TEST RESULTS ON:

$$Y_{r+1} \equiv (2^A + 1) Y_r + 1 \qquad (\text{mod } 2^{47})$$

$$Y_o = 2^{46}-1$$

TEST		A									THEOR VALUE
		5	10	15	20	25	30	35	40	45	
FREQUENCY X^2_9	1	7.8	11.0	9.0	8.0	900	900	45.8	5.2	165	$X^2_{9:05}$ = 16.919
	2	6.8	12.5	9.8	6.8	4444	103.9	6.8	3.5	1784	
SERIAL CORRELATION	1	-.092	.042	.073	.058	.999	.999	.912	-.113	-.350	0.00
	2	.081	.035	-.021	.023	.999	.971	.349	-.041	-.421	
MEAN	1	.510	.464	.469	.471	.250	.262	.435	.491	.418	0.500
	2	.503	.482	.493	.497	.289	.483	.497	.504	.424	
VARIANCE	1	.085	.086	.092	.081	.000	.000	.056	.076	.098	0.0833
	2	.084	.086	.084	.085	.001	.070	.081	.080	.100	
RUNS ≥ 6	1	1	1	1	1	> 15	> 15	> 15	2	0	.0285
RUNS ≥ 7	2	2	0	1	4	> 15	> 15	> 15	> 15	0	.0388

1. Sample size 100
2. Sample size 1000

TABLE 8.

TEST RESULTS ON:

$$Y_{r+1} \equiv (2^A + 1)\, Y_r + 1 \qquad (\mathrm{mod}\ 2^{47})$$

$$Y_o = 2^{46}-1$$

TEST		A									THEO. VALUE
		16	17	18	19	20	21	22	23	24	
FREQUENCY X^2_9	1	43.6	19.8	11.4	16.0	8.0	9.2	31.6	488	900	$X^2_{9:05}$ = 16.919
	2	8.0	12.7	3.7	12.6	6.8	18.9	272.1	4915	7477	
SERIAL CORRELATION	1	.648	.147	.068	.003	.058	-.255	-.049	-.404	.999	0.000
	2	-.008	.040	.003	.006	.023	.004	.017	-.398	.999	
MEAN	1	.448	.479	.469	.478	.471	.467	.435	.395	.250	0.500
	2	.503	.488	.497	.486	.497	.474	.413	.404	.270	
VARIANCE	1	.064	.073	.073	.075	.081	.082	.088	.051	.000	0.0833
	2	.079	.085	.085	.084	.085	.080	.081	.051	.000	
RUNS ≥ 6	1	4	2	2	1	1	0	0	0	>15	0.0285
RUNS ≥ 7	2	>15	4	0	7	4	0	0	0	>15	0.0388

1. Sample size 100
2. Sample size 1000

In applying the criterion of paragraph (1), the number of values of A reflecting a "poor" sample in the case of (3-2) is greater than for (3-1). Generator (3-2) is very susceptible to long runs and, above an A of 25, cannot be considered a practical random number generator. Using different starting values did not improve the output of (3-2); for $Y_0 = 1$ only an A = 21 passed and for $Y_0 = 2^{46} - 1$ no value of A passed. Even relaxing the criteria so that two failures are allowed before an A is considered "poor" only admits A = 18 and A = 20 as being acceptable for both starting values. Further, using a value of $C = (.788)\ 2^{47}$ as suggested by Coveyou [10] did not make any noticeable improvement in (3-2).

Table 9 shows the acceptable A's for generator (3-1).

TABLE 9

Yo	Acceptable A's								
1	10	15	18	19	21	22	35	40	
$2^{46}-1$	10	16	19	21	22	25	35	40	45

The above test results indicate that for these criteria generator (3-1) is much less erratic than (3-2). Six different values of A, (A = 10,19,21,22,35,40), are acceptable using both starting values and therefore it is suggested that A be chosen from these values.

Some tests were also made using the approach suggested by equation (2-1). Aside from some improvements in the poker tests results, no noticeable improvements were made. Using this approach increases the time of generation per word and unless the results registered a significant improvement, it could not be economically justified.

3. Results of Large Sample Tests.

Using the findings of paragraph (2) above, an A = 19 was selected for the five large samples to be tested. This particular value was selected over the faster A = 10 because it appeared to perform slightly better than A = 10 in the poker tests. The results are presented in Tables 10 through 17 and indicate that the samples satisfactorily pass the battery of tests. Figure 1 shows the average serial correlation of the five samples for lags one to ten. Over these lags the correlations ranged from +0.0182 to -0.0201.

4. Remarks on Computer Work.

All of the tests used in this paper were written using FORTRAN language. These have been catalogued and are available at the Computer Center at the United States Naval Postgraduate School. Programs using an A = 19 in Equation (3-1) have been written both for Scrap and for Fortran and are available in the computer library. Appendix B lists the machine language steps required to produce a random number according to (3-1) and the additional steps required to convert it to a floating point fraction between zero and one. These steps have been included to allow a programmer a little more flexibility in using the generator. In many cases, it may be easier to include these steps directly in a routine rather than go to a computer library subroutine.

5. Conclusions.

(1) To determine a suitable value of A in (3-1) the methods used in small sample testing are useful. Further refinements could be made by

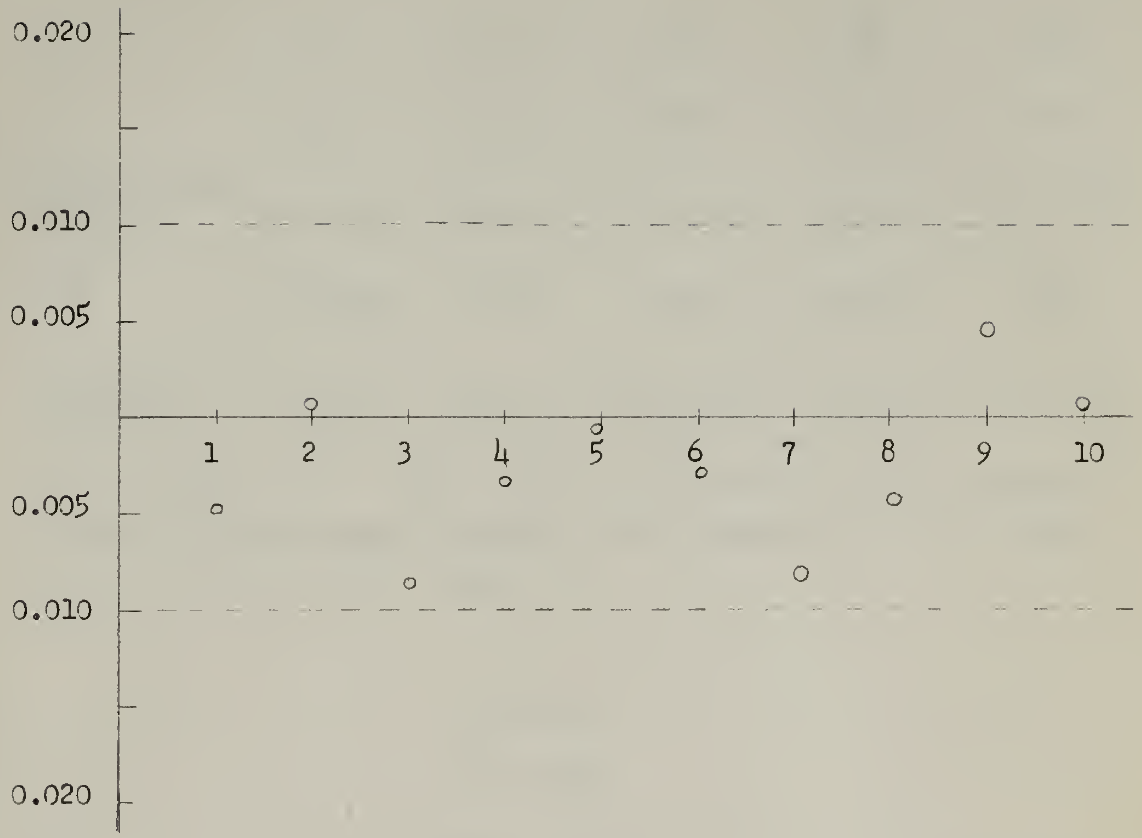

FIGURE 1.

AVERAGE SERIAL CORRELATIONS FOR TEN LAGS

TABLE 10

FREQUENCY TEST

	#1	#2	#3	#4	#5
X_9^2	3.39	11.66	5.46	13.36	7.41
PROBABILITY* INTERVAL	(.90,.95)	(.20,.30)	(.30,.70)	(.10,.20)	(.50,.70)
X_{99}^2	102.44	91.98	75.34	119.52	94.0
PROBABILITY INTERVAL	(.50,.25)	(.50,.75)	(.95,.975)	(.05,.10)	(.50,.75)

*(c,d) implies $c < P\left[X_v^2 > X_o^2 \right] < d$ where X_o^2 is the observed value of the chi-square statistic. For example, (.90,.95) implies $.90 < P\left[X^2 > 3.39 \right] < .95$.

TABLE 11

RUN TEST

RUN SIZE	NUMBER OBSERVED					EXPECTED NUMBER
	#1	#2	#3	#4	#5	
1	4105	4210	4166	4228	4104	4166.78
2	1806	1780	1822	1829	1851	1833.09
3	545	528	558	523	506	527.65
4	121	127	102	113	143	115.04
5 *	25	26	19	16	18	20.33
6	6	1	2	2	2	3.8

TABLE 12

RUN TEST CHI-SQUARE RESULTS

	#1	#2	#3	#4	#5
X_4^2	4.22	3.59	3.60	2.48	9.46
PROBABILITY INTERVAL	(.30,.50)	(.30,.50)	(.30,.50)	(.50,.70)	(.05,.10)

TABLE 13

POKER TEST (FIRST FIVE OCTAL DIGITS)

POKER HAND	#1	#2	#3	#4	#5	EXP. NUMBER
BUST	2004	2027	2000	2010	1979	2050.78
ONE PAIR	5073	5091	5109	5205	5163	5126.95
TWO PAIR	1623	1597	1578	1475	1616	1538.09
THREE OF A KIND	1035	1057	1050	1046	991	1025.39
FULL HOUSE	167	145	168	165	168	170.90
FOUR OF A KIND	96	81	92	98	83	85.45
FIVE OF A KIND	2	2	3	1	0	2.44

TABLE 14

POKER TEST CHI-SQUARE RESULTS (FIRST FIVE OCTAL DIGITS)

	#1	#2	#3	#4	#5
X^2_5	5.75	7.68	3.59	6.61	8.18
PROBABILITY INTERVAL	(.30,.50)	(.10,.20)	(.50,.70)	(.20,.30)	(.10,.20)

TABLE 15

CHARACTERISTICS OF GENERATED UNIFORM DISTRIBUTION
(FIVE SAMPLES OF 10,000)

	1	2	3	4	5	EXP. VALUE
MEAN	.50103	.50584	.50076	.50358	.50027	0.50000
VARIANCE	.08339	.084813	.08357	.08306	.08477	0.83333
SKEWNESS	.00029	.00073	.00014	.00002	.00001	0.00000
KURTOSIS	.01255	.01288	.01246	.01238	.012781	0.01250

TABLE 16

POKER TEST (SECOND FIVE OCTAL DIGIT POSITIONS)

POKER HAND	#1	#2	#3	#4	#5	EXP. NUMBER
BUST	2055	2053	2065	1997	2081	2050.78
ONE PAIR	5118	5104	5156	5148	5128	5126.95
TWO PAIR	1539	1560	1520	1561	1535	1538.09
THREE OF A KIND	1029	1008	1011	1032	995	1025.39
FULL HOUSE	168	192	166	160	167	170.90
FOUR OF A KIND*	89	79	75	101	89	85.45
FIVE OF A KIND	2	4	7	1	5	2.44

*Four and Five of a Kind combined for 5 d. of f.

TABLE 17

POKER TEST CHI-SQUARE RESULTS (SECOND FIVE OCTAL DIGIT POSITIONS)

	#1	#2	#3	#4	#5
X_5^2	0.18	0.982	1.32	4.85	1.86
PROBABILITY INTERVAL	>.99	(.98-.95)	(.95-.90)	(.50-.30)	(90-.80)

(a) selecting a few more starting values, Y_0, and tabulating the results as in Tables 1 through 4.

(b) modifying the criterion used in the sifting process to meet an individual's unique demands. For example, if he desired an unbiased sequence of uniforms, he could tighten the restriction on the acceptability limits of the mean.

(2) Based on the test results, Equation (3-1) is to be preferred to (3-2) as a generator of random numbers. The speed advantage in (3-2) of about 7.2 micro-seconds per word is masked by its demonstrated tendency toward long runs and poor frequency distributions for most values of A.

(3) Five blocks of 10,000 numbers each were tested using generator (3-1) with A = 19. All tests were passed satisfactorily and it is concluded that

$$Y_{r+1} \equiv (2^{19} + 3)\, Y_r \qquad (\text{mod } 2^{47}) \qquad (3\text{-}3)$$

will perform as a pseudo-random generator for the CDC 1604.

(4) The chaining of generators as suggested in Equation (2-1) did not lead to any positive indications that a better sequence of pseudo-random numbers would result. Unless this can be shown, the additional time of the generation is not warranted.

BIBLIOGRAPHY

1. M. G. Kendall and Babington Smith, Randomness and Random Sampling Numbers. Journal Royal Statistical Society, Volume 101, pp. 147-166, 1938.

2. The RAND Corporation, A Million Random Digits, The Free Press, Glencoe, Illinois, 1955.

3. D. H. Lehmer, Mathematical Methods in Large-Scale Computing Units. Second Symposium on Large-Scale Digital Calculating Machinery, at the Harvard University Computation Laboratory, Cambridge, Mass., Annals 26 (1951), pp. 141-146.

4. M. Greenberger, Appendix to Part II of "Decision-Unit Models and Simulation of the United States Economy", M.I.T., January 1958.

5. M. Greenberger, Notes on a New Pseudo-Random Number Generator, Journal of the Association for Computing Machinery, Volume 8, pp. 163-167, 1961.

6. A. Rotenberg, A New Pseudo-Random Number Generator, Journal of the Association for Computing Machinery, Volume 7, pp. 75-77, 1960.

7. M. Greenberger, An A Priori Determination of Serial Correlation in Computer Generated Random Number, Mathematics of Computation, Volume 15, Number 76, 1961.

8. H. B. Mann and A. Wald, On the Choice of the Number of Class Intervals in the Applications of the Chi-Square Test, Annals of Mathematical Statistics, Volume 13, Number 3, pp. 306-317, 1942.

9. M. G. Kendall, The Advanced Theory of Statistics, Volumes I and II, Griffin and Company, 1946.

10. International Business Machines Corporation, Reference Manual Random Number Generation and Testing, IBM, 1959.

11. R. R. Coveyou, Serial Correlation in the Generation of Pseudo-Random Numbers. Journal of the Association for Computing Machinery, Volume 7, Number I, pp. 72-74, 1960.

12. O. Taussky and J. Todd, Symposium on Monte Carlo Methods, John Wiley and Sons Inc., New York, 1956.

13. M. L. Juncosa, Random Number Generation on the BRL High Speed Computing Machines, Report Number 855, Ballistics Research Laboratories, Aberdeen Proving Ground, Maryland, 1953.

14. V. D. Barnett, The Behavior of Pseudo-Random Sequences Generated on Computers by the Multiplicative Congruential Method; Mathematics of Computation, Volume 16, 1962.

15. Control Data Corporation, Control Data 1604 Computer Programming Manual.

16. P. Peach, Bias in Pseudo-Random Numbers; Journal of the American Statistical Association, Volume 56, pp. 610-618, 1961.

17. T. Nagell, Introduction to Number Theory, John Wiley and Sons, Inc., New York, 1951.

APPENDIX A

GENERATOR CHARACTERISTICS

1. General.

In 1949, Lehmer [3] suggested using the multiplicative congruential method for generating pseudo-random numbers on a digital computer. Since that time, this method has been used, tested and reported on by a number of investigations, for example [12] [13] . In using this method on a binary computer of word size P, the recurrence relation most frequently takes the form

$$Y_{r+1} \equiv LY_r \qquad (\text{mod } 2^P) \qquad \text{(A-1)}$$

where L is a fixed odd integer and $0 < Y_r \leq 2^P - 1$ being relatively prime to 2^P. This sequence is periodic, its length depending on the choice of L, Y_r and P. When the indicated iteration is performed, say n times, (1-1) becomes:

$$Y_{r+n} \equiv (L)^n \ Y_r \qquad (\text{mod } 2^P)$$

and for this reason is sometimes referred to as the power residue method.

The numbers $Yr/_2P$ (r = 0,1,2....) become the desired uniformly distributed elements from the unit interval. Barnett [14] in 1961 showed that necessary and sufficient conditions to obtain a maximum cycle of 2^{P-2} elements, all of which are distinct, are:

(a) $L \equiv 3 \pmod 8$ or $L \equiv 5 \pmod 8$ (A-2)

(b) $Y_0 \equiv 1 \pmod 2$, (A-3)

that is, Y_0 must be odd.

If these conditions are not met the cycle length can be affected and in some cases the process degenerates to zero.

2. Discussion of Equation (1-2).

While the sequence produced by (1-1) has been found satisfactory, nevertheless, it does imply the use of the multiplication instruction in the computer. Relatively speaking, this is a time-consuming operation and interest continued on methods that would avoid its use and still produce the desired sequence. Greenberger [4] proposed using an L of the form $(2^A + 3)$ where $2 \leq A < P$. It is seen that such an L satisfies the conditions necessary to insure a cycle of 2^{P-2}. This form has the advantage of substituting a shift operation of size A and three add instructions for the multiplication operation. In the CDC 1604, the multiplication instruction [15] requires 25.2 plus .8n micro-sounds; where n is the number of ones in the multiplier. On the other hand, the shift instruction takes 2.8 plus .4s micro-sounds, where s is the number of places shifted and an add instruction takes 7.2 micro-sounds. In operations where large quantities of random numbers are required this saving can make a significant contribution toward reducing machine time.

When this form of L is translated to the CDC 1604, equation (1-1) becomes:

$$Y_{r+1} \equiv (2^A + 3) Y_r \qquad (\text{mod } 2^{47}) \qquad \text{(A-2)}$$

where with $2 \leq A < 47$ the period of the sequence is 2^{45}, that is, 2^{45} distinct numbers are produced before the sequence repeats. Each of the numbers produced are odd and belongs to one of two mutually exclusive integer sets depending on the selection of Y_0 where, if r belongs to one set r+2 cannot. These two sets between them exhaust all of the possible odd numbers in the interval 1 to 2^{47}. Because the numbers are odd this means that the two least significant bit positions in the binary representation of a number will not change, that is, their cycle is zero. The third bit position has period 2^1, the fourth 2^2 and so on. For this reason, if random digits are to be used only the most significant digits of a number should be considered.

No mathematical rules have been developed to accurately determine a particular choice of A although it is clear that an unfortunate selection here can result in an unacceptable generator. Coveyou [11] did offer a criterion for selecting A in terms of an approximation formula he derived for the reduction of serial correlation between the numbers. Subsequently, Greenberger [5] introduced a correction term to be applied to Coveyou's expression and demonstrated that even though serial correlation might be optimized, nonetheless, the generator might suffer from other shortcomings, such as producing long sequences of low numbers.

3. Discussion of Equation (1-3)

Rotenberg [6] tested a sequence

$$Y_{r+1} \equiv L\, Y_r + C \qquad (\text{mod } 2^P) \qquad (A-4)$$

where L and C are odd integers less than P and $0 \leq Yr \leq 2^P$. In this case, a maximum period of 2P is obtained for any

$$L \equiv 1 \qquad (\text{mod } 4) \qquad (A-5)$$

An L of the form $(2^A + 1)$ with $A \leq 2$ satisfies the condition of (A-5) and was used by Rotenberg in his tests. He examined two values of C, C = 1 and C = $(.788)\ 2^P$, the latter value resulted from Coveyou's approximation mentioned above. Rotenberg reported that there were no significant differences in his results using either one of these C values.

For the CDC 1604, this generator becomes

$$Y_{r+1} \equiv (2^A + 1) Y_r + C \qquad (\text{mod } 2^{47}) \qquad (A-6)$$

This type has an edge on form (A-2) in that it requires one less add instructor for each new number. It also has a longer period although this does not seem too significant in that (A-6) will produce 2^{45} distinct elements.

4. Sub Periods.

It has been pointed out by Peach [16] that in the case of (A-6) certain harmonics exist throughout its period. These tend, he feels, to lend a greater stability to the generated numbers than should be expected of a random sample. He illustrates his point by making frequency tests on large samples and notes that they appear to be unduly uniform, although his results do not overwhelmingly indicate this. Whether too much stability is objectionable depends on the particular application. If a user wanted to estimate an unknown mean without bias then this characteristic might not be critical. If the variance is a point of interest too much stability could be critical.

5. An Example.

To illustrate the multiplicative congruential method consider equations (A-4) which with L = 5, $Y_o = 0$, C = 5, P = 3 becomes

$$Y_{r+1} \equiv 5 (Y_r + 1) \qquad (\text{mod } 2^3) \qquad (A-7)$$

The sequence is determined as follows.

Substituting $Y_o = 0$ into equation (A-7) gives

$$Y_1 \equiv 5 \qquad (\text{mod } 2^3)$$

This means that Y_1 is the remainder after dividing 5 by 2^3. This gives $Y_1 = 5$. Substituting this back into (A-7) gives:

$$Y_2 \equiv 5\ (6) \qquad\qquad (\text{mod } 2^3)$$

Dividing 30 by 2^3 gives a remainder 6. Therefore $Y_2 = 6$. This is repeated giving a sequence $Y_1 = 5$, $Y_2 = 6$, $Y_3 = 3$, $Y_4 = 4$, $Y_5 = 1$, $Y_6 = 2$, $Y_7 = 7$, $Y_8 = 0$. of period 2^3.

APPENDIX B

PROGRAMMING THE GENERATOR

Method

1. The Relation.

$$Y_{r+1} \equiv (2^{19} + 3) \; Y_r \pmod{2^{47}}$$

$$Y_o = 2^{46} - 1$$

can be programmed for the CDC 1604 as follows:

(a) Define:

```
MASK  = 4000000000000000 Binary
MASK2 = 2000000000000000 Binary
```

$$Y_o = 2^{46} - 1$$

MACHINE LANGUAGE ROUTINE:

```
ENQ (0)
LDA (Xo)
LLS (19)
SCL (MASK)
ADD (Xo)
ADD (Xo)
ADD (Xo)
STA (Xo)
```

(b) At this point, the random number is in the accumulator and has been stored in X_o in preparation for the next number if required. It is now required to convert it from its present integer form to a floating point fraction on the unit interval. This is done by continuing on with the following instructions:

```
ARS (11)
ADD (MASK2)
FAD (MASK2)
```

At this time, the desired uniformly distributed 0,1 number is in the A- register. The time of generation is about 90 micro-seconds per number.